DOCUMENT RESUME

ED 282 414 FL 016 705

AUTHOR de Jong, John H. A. L.
TITLE Tailoring Tests to Educational Levels.
PUB DATE 84
NOTE 15p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Difficulty Level; *Efficiency; Equated Scores; Foreign Countries; Language Proficiency; *Language Tests; *Listening Comprehension; Secondary Education; Second Languages; *Standardized Tests; Statistical Analysis; Test Length; *Test Reliability
IDENTIFIERS *Netherlands

ABSTRACT

The Netherlands' secondary education system is highly differentiated, with four different school types for four scholastic ability levels. Final examinations must accommodate these four levels, and require a test-independent definition of the intended final ability levels as well as a sample-free evaluation of the range of ability levels at which a particular test will measure with sufficient accuracy. A method for locating, on a single scale, the ability distribution of a given population as well as the test's optimal reliability level is proposed and illustrated. The method is demonstrated using standardized tests of foreign language listening comprehension for the two highest levels of secondary education in the Netherlands. The languages tested are French, German, and English. Statistical analyses of the test results for each group are presented and discussed, and implications for test use are examined. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED282414

John H.A.L. de Jong
Cito, National Institute for Educational Measurement
Arnhem, Netherlands

TAILORING TESTS TO EDUCATIONAL LEVELS

1 Introduction

Secondary education in the Netherlands is a highly differentiated system. Four different schooltypes aim at four different scholastic ability levels. (Fig.1) In such a system the final examinations must match these different levels and they therefore require a test-independent definition of the intended final ability levels as well as a sample-free evaluation of the range of ability levels at which a particular test will measure with sufficient accuracy. The present paper presents a method to locate on a single scale the ability distribution of a given population as well as the level at which a particular test will yield optimally reliable estimations of the ability of a person taking the test. The method can therefore be used to evaluate the fitness of a test for pass/fail dicisions in a particular population and, in the occasion, to adapt pretests to that level.

The fitness of a test is a subjective argument as it is determined by weighing reliability against efficiency. The discriminatory power required of a test is a function of the range in ability within the group to be tested and the importance of decisions based on the results. In the Netherlands high discrimination power is demanded as the ability continuum present in secondary education is subdivided in different schooltypes whereas within schools grade marks are distributed on a ten point scale (which in fact is a hundred point scale as marks are given to one decimal) and pass/fail decisions are based on these marks. To achieve high discriminatory power at a certain ability level an item has to be of the appropriate difficulty. Any item that is too difficult or too easy will not discriminate optimally. To achieve reliability of test results at a certain ability level, a number of items of the appropriate difficulty is required. To assess the ability of persons in a group varying in ability a number of items of varying difficulties is

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

de Jong

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

FL016705

BEST COPY AVAILABLE

necessary. If variation in ability is high, tests will become too long. In such circumstances a series of tests of varying difficulty is more efficient. The importance of correct level assignment of tests can be seen in figures 2, 3 and 4. Figure 2 presents the item characteristic curve (I.C.C.) for any item. The sum of the I.C.C.'s of all items in a test yields the test characteristic curve (T.C.C.) which theoretically has the same shape. The point is that each item, and each test have highest discrimatory power in one particular region of the ability continuum only and will lose discriminatory power in groups exhibiting either too low or too high ability. This phenomenon is often referred to as floor, and ceiling effects. Figure 3 presents the T.C.C. of a pretest actually passed in two groups of different ability levels. The test clearly discriminates better in group A which means it results will be more accurate, reliable and significant. The reverse case is presented in figure 4, here the effect on KR20 reliability is even more dramatic.

The method proposed to asses fitness will be demonstrated using Cito tests of foreign language listening comprehension for the two highest levels of secondary education in the Netherlands.

2 Material

Cito test of foreign language listening comprehension have been constructed at Cito and used in final examinations in schools for over ten years. The tests were originally developed in a research project at the University of Utrecht (Groot, 1975). The objective of the tests, as defined by Groot, is to evaluate foreign language learners' ability to understand the foreign language, spoken spontaneously by educated native speakers in normal conversational tempo. Extremely informal elements as well as lexical and syntactical elements that are incomprehensible to less educated native speakers and topics requiring specific knowledge are excluded from the language material in the tests. The language material consists of three or four interviews with native speakers of various occupations and professions and, for the highest level, a lecture on a subject of general interest is included in the tests. The item format is a multiple choice question with three options. The language material is divided in samples of fourty to fifty seconds; the correct answer is a one phrase summary of the global contents of the sample. A pause of twenty to twentyfive

seconds is provided on the tape in between samples to allow for the testees to read the item in their test booklet and tick their anser on an answer sheet. Test length varies from 40 to 50 items depending on school level. Administration time varies accordingly between 50 and 60 minutes. Test reliability (KR20) ranges from .70 to .85 with a noticeable tendency towards the higher values in the last few years. Acceptability of the tests is clearly reflected by the fact that more than 85 percent of the Dutch secondary schools use them in their final examinations. Though this description applies to the tests used in the present research design, since 1981 the scope of the tests has been enlarged by using different item formats e.g., modified cloze items allowing for a wider sample of language material in the tests (De Jong, 1983).

The two highest levels of secondary education in the Netherlands are HAVO and VWO. HAVO is an intermediate level of general (nonvocational) education, meant as a basis for further education preparing for higher non-academic functions. HAVO is a five year programme. Foreign languages are taught at an average rate of 3½ weekly periods. The final level is comparable to that obtained in most western countries at the secondary school level.
VWO, the highest level of general education, is meant as a preparation for further academic studies. VWO is a six year programme. Foreign languages are taught at an average rate of three periods a week. (Fig. 1)

The ability level required to understand a given sample of language material is hard to define as it is related to a complex of factors such as: range and choice of vocabulary, grammatical complexity, articulation and other speech features of the speaker, abstraction level of the message, density of information and acoustical conditions. Though a number of these factors can be measured, their effect on the difficulty of language material, and most certainly the interrelated pattern of influences of different factors, present in varying degrees, remains uncertain. There is no satisfactory solution to the measurment of language difficulty but the holistic approach. In practice test constructors have to rely on their intuition and the advice of a board of teachers in order to match samples of language and educational levels. A pretesting procedure including the assessment of fitness of tests for the level aimed at would constitute a holistic but reliable check on the intuition

based level assignment with the further advantage of the possibility to adapt the test to the required level by a careful selection of items.

3 Method

The method is based on item response theory using the Rasch model (Rasch, 1960). In the Rasch model one item parameter and one person parameter determine the probability that a given subject aswers a given item correctly. The item parameter in the Rasch model is called the difficulty parameter and the person parameter is called the ability parameter. However, since for the calibration of each test an arbitrary origin of measurement is selected the numerical values of the parameters of different tests cannot be compared directly.

In order to make them comparable the same origin has to be chosen for the different tests. This procedure, known as test equating, can be followed if the tests measure the same ability and if one of the following conditions is met:

- the tests contain a number of common items;
- the groups of persons taking the tests contain common persons;
- the groups of persons taking the tests are representative samples from the same population.

If both conditions are met and the different tests do not differ to much in difficulty (Petersen e.a., 1982) the item parameters of the tests and the ability parameters of persons taking the tests can be equated.

Test equating in this study was done with a number of common items: Havo-pupils made part of a VWO test and VWO-pupils made part of a Havo test apart from each group making its own test (fig. 5). The items made by both groups constitute a so-called anchor test. After seperate calibrations of each test (including the anchor test) a pair of independently estimated parameters of each anchor item was available. As a result of sample independence of the estimation, these parameters are equivalent except for an additive value that, apart from sampling errors, is the same for all pairs in the anchor test. Once this additive constant on the anchor test has been calculated, by taking the mean of the differences within all pairs, it can be added to the item parameters of all items in one test to bring the test on the same scale as the other test in the equating procedure.

Test analysis was done with the computer programme CALFIT (Wright and Mead, 1975).

The fit of the anchor-items was evaluated by

$$(\sigma_{iN} - \sigma_{iV} - C_{H,V})^2 (N/12)[K/(K-1)], \qquad (1)$$

which is approximately χ^2-distributed with 1 df (Wright and Stone, 1979). Total anchor fit was evaluated by

$$\sum_{i}^{K} (\sigma_{iH} - \sigma_{iV} - C_{H,V})^2 (N/12)[K/(K-1)] \sim \chi^2_K, \qquad (2)$$

where

σ_{iH}, σ_{iV} = parameter estimates for item i in the anchor for the HAVO and VWO test, repectively

$C_{H,V}$ = $\sum_{i}^{K} (\sigma_{iH} - \sigma_{iV})/K$, the mean of difference within parameter pairs

K = number of anchor items, and

N = number of pupils presented with the anchor part of the test.

After bringing both sets of parameters on a common scale, the amount of information yielded by either test and thus the accuracy with which it measures at any particular ability level can be calculated with the formula

$$I = \sum_{j}^{k} \frac{\exp(\xi_i - \sigma_j)}{(1+\exp(\xi_i - \sigma_j))^2} \qquad (3)$$

where ξ_i is the ability parameter for subject i and σ_j is the difficulty parameter for item j, k the number of items.
The higher the information function, the more reliably decisions can bemade. The standard error of measurement at a given ability level can be calculated directly from the information function with

$$S_e = \frac{1}{\sqrt{I}} \qquad (4)$$

In order to assess the fitness of a test for a particular schooltype the distribution of ability within that schooltype has to be known. The information function reveals the area where the test measures best, the ability distribution reveals whether this area is important in the schooltype in question.

Of course, purposes of testing can vary widely. The aim might be, for example, to select the highest one percent of the distribution in order to award scholarships or else to select the lowest 25 percent, to deny entrance. In fact for the tests under investigation here the primary purpose is to award qualifications as 'sufficient' or 'insufficient',: pass or fail.

The pass-fail borderline, or cut-off point, generally lies somewhat to the left of the modus of the distribution. It seems natural to demand that accuracy of measurement is highest in that region, where these most important decisions are made.

In view of the reasons mentioned above, the choice of items should be such, that the sum of their information functions, i.e. the test information function, reaches an optimum in the desired region of the ability distribution. With Rasch-calibrated items the maximum is N x 0.25, where N is the number of items. This maximum is reached when for a unique ability level, all items have a difficulty that equals exactly the chosen ability level. Since in that case (3) becomes N x 0.25. It is in this sense, that a test should consist of items best suited for a particular level.

The distribution of true ability within the HAVO and VWO group has been calculated with a computer programme developed by Verstralen (1982). In general practice, as in this case too, the distribution of true ability does not differ much from the observed distribution of ability, provided tests are sufficiently reliable and samples are sufficiently large. In this paper therefore the observed distribution is used whenever it is more convenient.

Results and discussion

Figure 6 presents the information function of a pretest of French listening comprehension meant for the VWO level together with an indication of the observed distribution of ability in the VWO group. The pretest consists of 75 items which is, deliberately, too long for use under examination conditions. A

selection of the most appropriate subset of 50 items will have to be made. The pretest is off target as the information function reaches its maximum in an area where VWO pupils are not represented. The average random selection of 50 items from the 75 pretested items would reduce test length only, without improving the information function. By a careful selection, however, of items at the difficulty level corresponding to the ability level at the cut-off point it is possible to achieve maximum information exactly in the relevant area where pass/fail decisions will be made.
The selected items constitute a test considerably shorter than the original pretest whereas it yields almost the same amount of information for determining the ability of most pupils in the target group.

Figure 7 presents the information function of the 50 items selected for the actual tests of French listening comprehension at the HAVO and the VWO level on the same scale, together with the estimated distributions of true ability for both populations.
The HAVO and VWO population are sufficiently different to merit seperate tests. Items e.g., which the modal VWO pupil will pass in about 50% of the times will be answered correctly in less than 30% of occurences by the model HAVO pupil. On the other hand the difference between both populations is not too large to endanger the equating procedure. There is less than one z-score difference between the means of both populations on either test. The maxima of the test information functions of both levels practically coincide with the region where pass/fail decisions will be made.

Figure 8 presents indications of the observed distributions of ability in HAVO and VWO schools for French, German and English listening comprehension. On the ability continuum is shown how the observerd ability is transformed into grade marks at both levels. In the Netherlands the conversion of scores - i.e. observed ability - in grade marks is still calcutaled in relation to test results. The cut-off point is chosen at roughly 20% of the cumulative frequency and assigned 5.5, maximum score is always 10. This implies that, whatever the range in ability within a certain schooltype is, marks will always be distributed from 1.0 to 10.0.
If then the range is small, as is the case for French, the difference between pupils earning low and high marks will also be small, whereas if the range is larger, as for English, it takes a greater difference in ability to obtain a

higher mark. The range in French is probably smaller because only 30 to 40% of the pupils take French as subject whereas English is taken by over 90% of the population. German takes an intermediate position with 40 to 50%. Another effect of the wider distribution in both HAVO en VWO English is the lager amount of overlap in ability between both populations, which in its turn causes a smaller difference between marks given at the HAVO and the VWO level. The average HAVO pupil would have a fair chance to pass an English VWO test but he would certainly fail in French.

It would only be fair to let HAVO pupils present themselves at VWO exams in English. However, as long as the cut-off point is determined in relation to test results this would necessarily entail lowering VWO Standards. It would therefore be recommendable that the relative determination of cut-off points be subtituted by an absolute method. By equating tests of subsequent years and choosing a fixed cut-off point on the ability scale the standard for passing would be made independent from the group of pupils going in for an examination. It would seem reasonable to take the average cut-off point of a number of recent years as the new absolute cut-off point.

Furthermore it should be equally discussed whether the distinction in schooltypes holds for all subjects in the same way. Given the much wider distribution of ability in VWO English than in VWO French one could e.g. consider distinguishing two seperate levels within the VWO population for English. The determination and evaluation of school curricula is ultimately a political issue, new methods in test analysis can offer sound criteria for decisions at the political level.

References

Groot, P.J.M., Testing communicative competence in listerning comprehension; in: R.L. Jones and B. Spolsky, (eds.), *Testing language proficiency*; Center for Applied Linguistics, Arlington, Virginia, 1975.

Jong, J.H.A.L. de , Focusing in on a latent trait: An attempt at construct validation by means of the Rasch model; in: J. van Weeren, (ed.), *Practice and problems in language testing 5: Non-classical test theory. final examinations in secondary schools;* National Institute for Educational Measurement, Cito, Arnhem, 1983.

Petersen, N.S., G.L. Marco and E.E. Stewart, A test of the adequacy of linear score equating models; in: P.W. Holland and D.B. Rubin, (eds.), Test equating; Academic Press, New York, 1982.

Rasch, G., *Probabilistic models for some intelligence and attainment tests,* Danmarks Paedagogiske Institut, Copenhagen, 1960.

Verstralen, H.H.F.M., *WAVSIM en VERPLT; twee programma's voor de normering van een Rasch gecalibreerde itembank*; Cito, Arnhem, 1982.

Wright, B.D. and R.J. Mead, *CALFIT*: Research memorandum number 18; Department of Education, University of Chicago, Chicago, 1975.

Wright, B.D. and M.H. Stone, *Best Test Design*; Mesa Press, Chicago, 1979.

Fig 1 Secondary schools in the Dutch Educational system

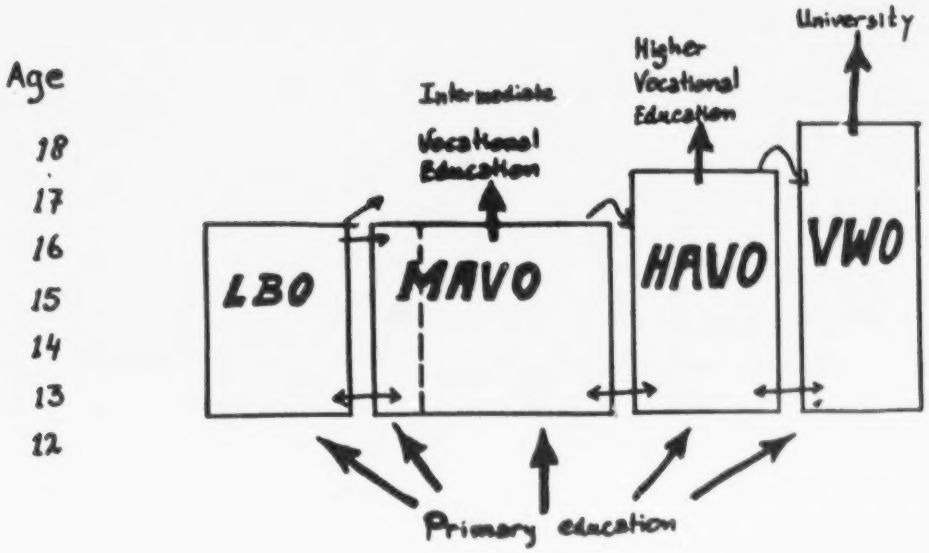


Fig 2 Item characteristic curve / Test characteristic curve

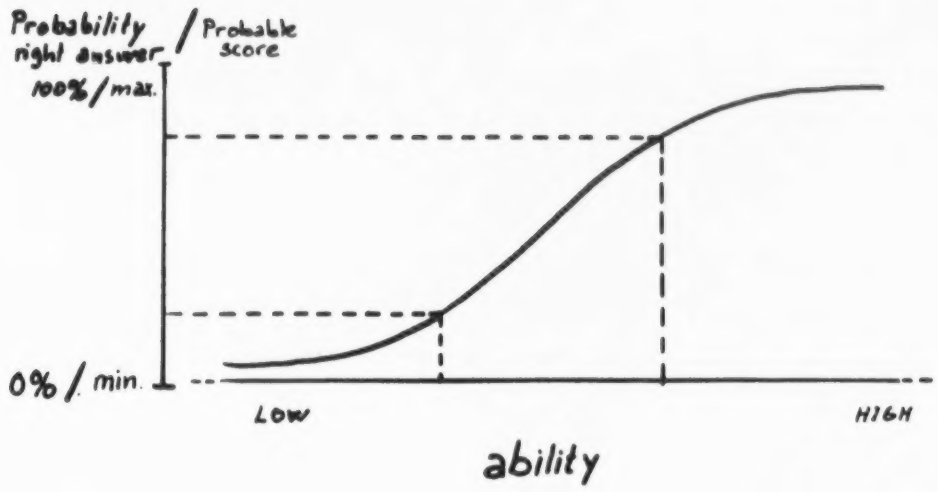
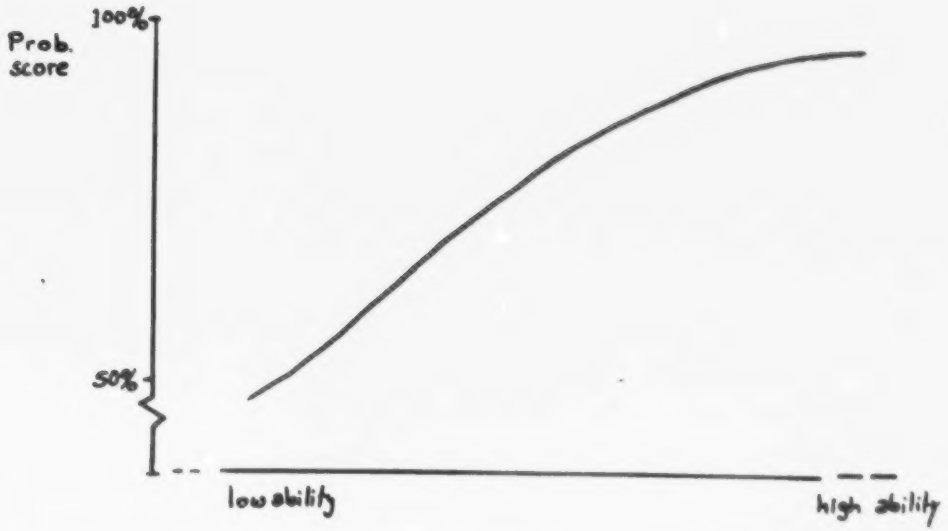
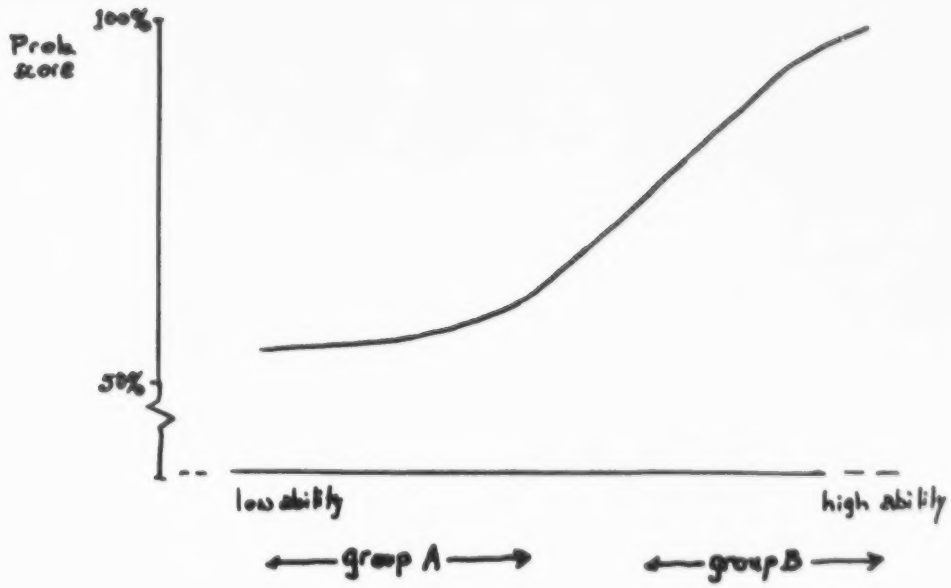


Fig 3 Test characteristic curve for test x, too easy for group B



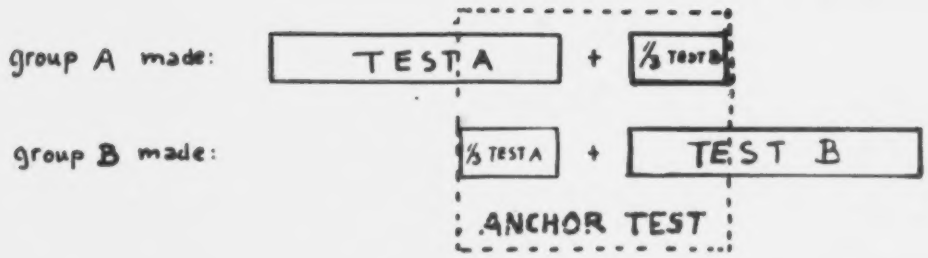
KR 20 .75 .68

Fig. 4 Test characteristic curve for test y, too difficult for group A



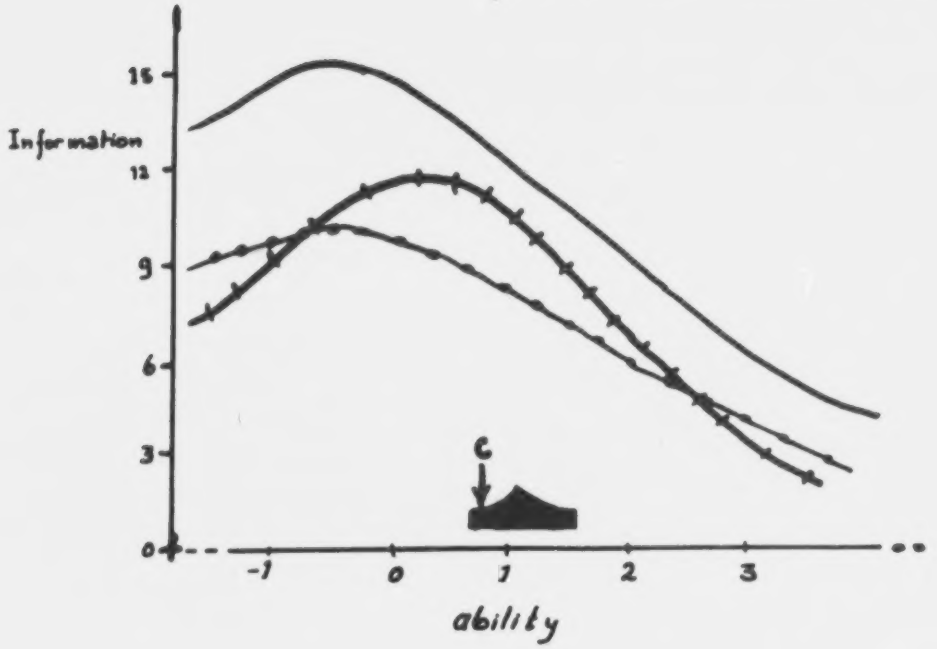
KR 20 .07 .71

Fig. 5 Design of equating procedure



number of items	in original tests	: 75
	in anchor test	: 50
	answered by each pupil	: 100
number of pupils in samples		: 250

Fig 6 Information function French VWO — Pretest 75 items
-o-o-o- random selection 50 items
-+-+- test selection 50 items



68% of testees
cut-off point

Fig 7 Test information functions and population distributions HAVO and VWO French listening comprehension

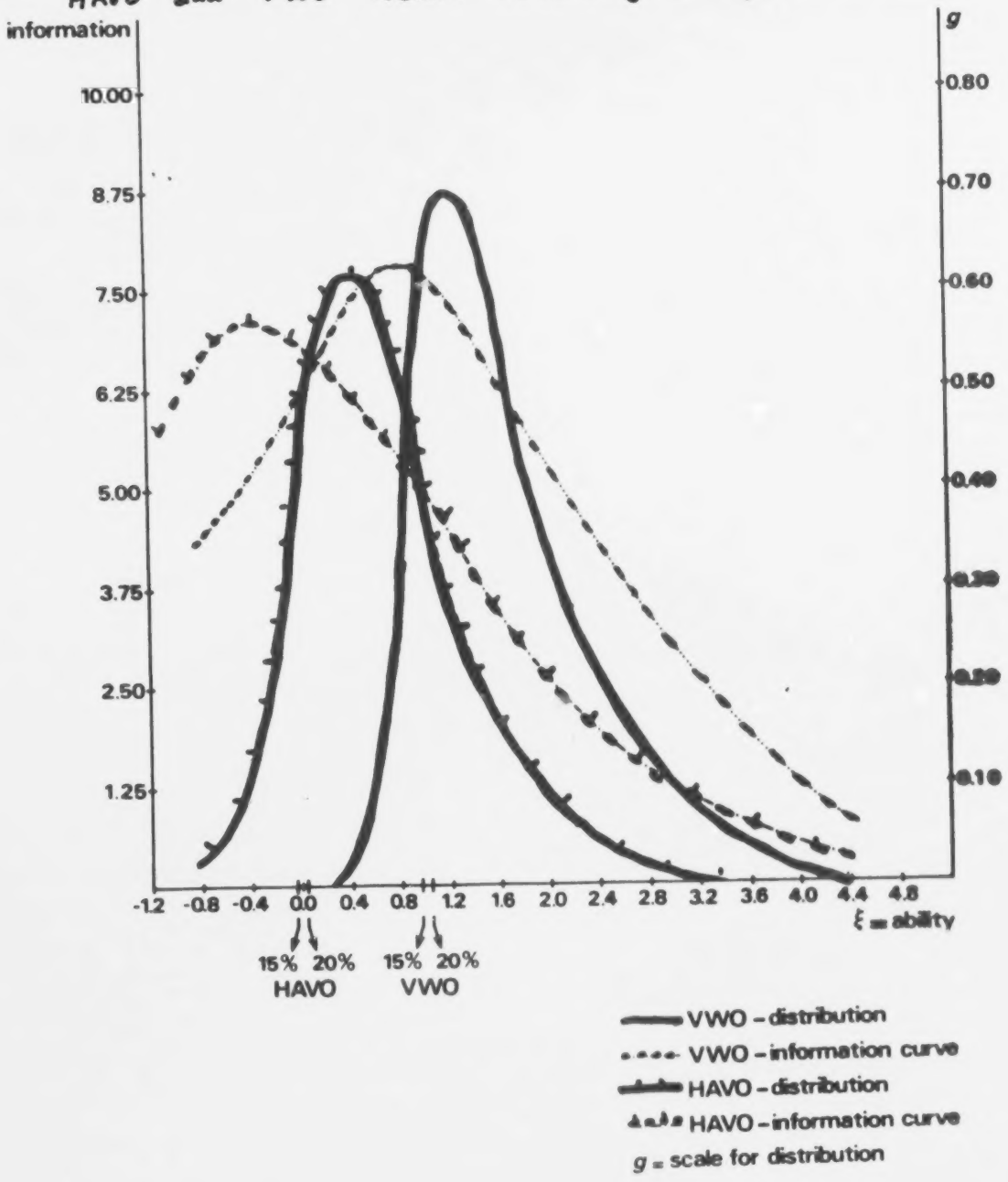
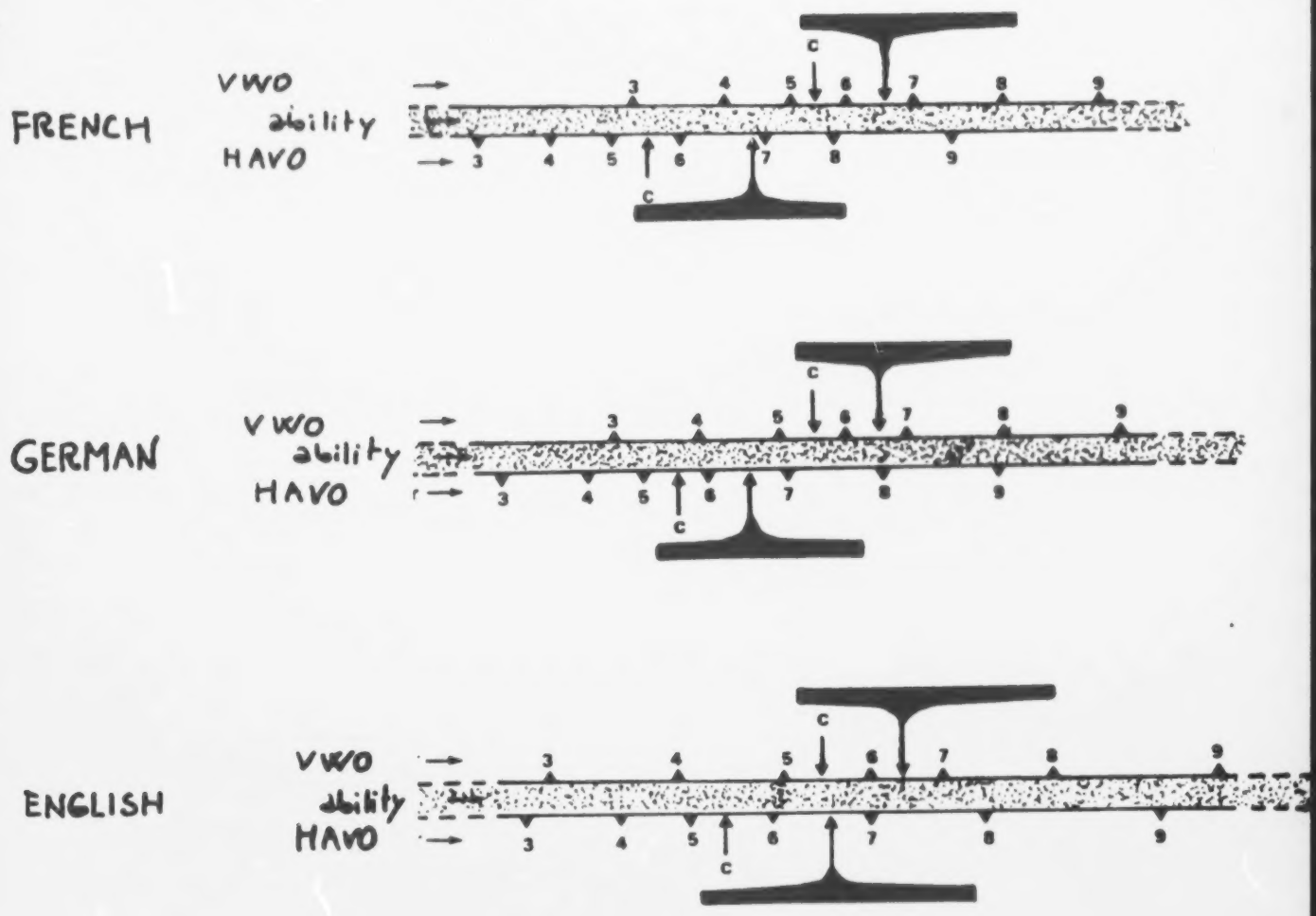


Fig. 8 Ability, grade marking and distribution in HAVO and VWO populations for French, German and English



C : passing score / cut-off point

: mean (point of arrow) and 68% of distribution